NRL Report 7773

Quality and Acceptability Testing of Voice Processors for Military Applications

DAVID C. COULTER

Systems Integration and Instrumentation Branch
Communications Sciences Division

November 15, 1974

NAVAL RESEARCH LABORATORY
Washington, D.C.

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: NRL Report 7773	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE *(and Subtitle)*: QUALITY AND ACCEPTABILITY TESTING OF VOICE PROCESSORS FOR MILITARY APPLICATIONS	5. TYPE OF REPORT & PERIOD COVERED: Initial report on voice quality and acceptability testing.
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): David C. Coulter	8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Naval Research Laboratory, Washington, D.C. 20375	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: NRL Problem 54R01-62, X3299, Task 100, 200, 300, 400
11. CONTROLLING OFFICE NAME AND ADDRESS: Naval Electronic Systems Command, Department of the Navy, Washington, D.C. 20375	12. REPORT DATE: November 15, 1974
	13. NUMBER OF PAGES: 22
14. MONITORING AGENCY NAME & ADDRESS *(if different from Controlling Office)*	15. SECURITY CLASS. *(of this report)*: Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE
16. DISTRIBUTION STATEMENT *(of this Report)*	
17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*	
18. SUPPLEMENTARY NOTES	

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

Quality and acceptability testing	Vocoder/PSK
Toll Quality	Analog voice PDM/PSK
Gaussian noise background	Variable slope delta modulation/PSK
Confidence interval	Analog voice FM
Rank ordering	Digital voice PDM

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

The application of digital voice communication systems for military usage has been hampered by the lack of reliable predictions of system voice quality and acceptability. Although intelligibility tests have proved useful to system designers and have provided lower limits to system acceptability, there has been no simple method for rank ordering voice processors operating in a given transmission environment with respect to voice quality. A method for rank ordering which is useful for up to about ten systems and gives reliable results is described in this report

(Continued)

UNCLASSIFIED

20. Abstract (Cont.)

along with a typical system evaluation at NRL. This evaluation shows the ability of the test to rank order systems under simulated transmission conditions. It should be noted that adding a new system to a previous ranking with this method is difficult and may require rerunning the entire test.

CONTENTS

QUALITY AND ACCEPTABILITY TESTING OF VOICE PROCESSORS FOR MILITARY APPLICATIONS

INTRODUCTION

For many military applications of voice communications the available circuits, such as wireline, satellite and HF, do not have the capacity for the data rates normally used by digital voice systems including continuously variable slope delta modulation (CVSD): 16 kbps to 32 kbps; pulse code modulation (PCM): 48 kbps to 56 kbps. To reduce the data rates, a narrowband digital voice processor must be employed.

Although narrowband digital processors are usually capable of providing voice transmission at reduced data rates with usable intelligibility, they often produce a degraded-quality voice signal at the receiver. The concern induced by these systems is centered mainly on the quality of the received synthesized speech. In this sense the term *quality* refers mainly to the factors influencing user acceptance for the purpose of either Navy or other military communications and does not include all the factors implied by the Bell Telephone System in *toll quality*, such as speaker recognition.

BELL TELEPHONE SYSTEM VOICE TRANSMISSION

In the Bell Telephone System's voice communication channels the goal is to provide *toll quality* voice links for the users which implies a certain degree of customer acceptance. The definition for *toll quality* is based on the achievement of certain factors such as absence of distortion, good signal-to-noise ratio, and adequate frequency response.* Combined, these factors yield a system that is found to be acceptable by a vast majority of users. Additionally, the term *toll quality* implies a considerable capability for speaker recognition, once the speaker's telephone voice has been learned (once it is learned how the distortions of a typical telephone circuit affect a particular voice).

Voice reproduction which meets the *toll quality* criteria at data rates substantially lower than those normally encounted in the use of PCM equipment has been the subject of ongoing research at the Bell Telephone Laboratories for many years. However they have not yet successfully implemented a low-bit-rate voice digitizer which produces *toll*

*Unfortunately, these factors are impossible to measure directly in a system using narrowband digital voice processors, such as the vocoder and linear prediction encoder. These systems are based on separate voice pitch and spectrum transmission and will not respond to normal measurement techniques such as the signal-to-noise ratio in various frequency bands or sinewave signal transmission. Such measurements, if attempted, could give grossly misleading results. It is for this reason that direct voice measurements such as intelligibility and quality must be made, which can then be compared with similar ratings for analog or PCM circuits.

Manuscript submitted May 28, 1974.

quality speech. Even if this processor became a reality, the problem of its cost versus those of existing high-data-rate digital voice equipment and the potential tradeoff in reduced circuit costs would remain.

MILITARY-CIRCUIT VOICE TRANSMISSION

The constraint of *toll quality* on voice communication usually does not apply to military communications. Consumer public-relations considerations such as received voice naturalness are areas over which military necessity, in theory, should take precedence. In addition, military requirements for speaker recognition (voice authentication) can be more accurately performed by electronic means rather than by relying on auditory perception.

Since a small but growing set of applications absolutely requires voice communication at low bit rates, a number of systems employing available vocoders operating at 2400 bps have been designed and deployed. Although the overall satisfaction has not been overwhelming, the systems with vocoders have unquestionably met crucial needs. Since user satisfaction has not been complete, this report will further the efforts toward choosing a system having more universal acceptance. It will describe a method used by the Naval Research Laboratory for measuring the relative *acceptability* of such systems, as an aid to selecting future systems with improved voice quality. (The terms *quality* and *acceptability* are defined in this report as relative terms. If a system could be found whose quality was such as to be just marginally acceptable, then systems ranked lower could be called *unacceptable* and systems ranked higher, *acceptable.)*

NEED FOR LISTENER QUALITY AND ACCEPTABILITY TESTING

Many candidates are being proposed for low-bit-rate voice processors for military applications; hence to aid system procurement some method is needed to determine the overall acceptability of system voice quality. System voice quality should be a factor in system selection in conjunction with factors such as intelligibility, equipment complexity, reliability, life-cycle cost, and ease of operation.

Despite years of intense research in speech reproduction by narrowband voice processors, the state of the art is still far from perfection. The results of present testing methods, such as intelligibility testing, have not been completely representative of actual system acceptability in the field. For this reason some persons have advocated no formal testing, or only conversational tests by experienced listeners, for determining voice quality and acceptability. Although such tests are highly recommended to detect problem areas which might possibly be missed by other tests and to give overall *suitability* judgments, they cannot possibly be substituted for more formal acceptability tests. Rather, it has become important to continue improving formal tests based on improved characterizations of the factors relevant to good voice reproduction.

Any further progress in the state of the art of system design, as well as intelligent choices between present systems, must be based on statistically representative and repeatable tests. Several factors have led to a general lack of faith in such tests in the past:

- Intelligibility test results have not always predicted user acceptance. There appears to be some lack of correlation between voice intelligibility and acceptable voice quality.

- Past test efforts have involved too few samples of different voices to be representative of the general population. Some systems were designed to favor only the voice of the design engineer, and he in turn learned to talk in a way to favor the best system performance. Thus, sometimes a vocoder was found acceptable based on this one individual, even though the test results for many talkers would most likely indicate a system which is unacceptable compared with other equipments designed for the same general usage. Therefore with respect to system design it seems important to evaluate the question of voice sensitivity.

- All current intelligibility[1-3] and quality[4-7] tests make use of human listeners as the test instrument (as do the informal conversational evaluations). To get reliable test results from these tests, the listener's task must be simplified as much as possible: For example, asking the listener to compare different systems using identical sentence material for each system makes judgment easier.

- Past test comparisons have stressed optimized performance (use of good talkers, trained listeners, well adjusted equipment, no competing tasks, etc.), whereas actual conditions have always represented some degradation from these ideals. This has led to informal subjective evaluation, generally consisting of a judgment, hopefully unbiased by system loyalties, of how difficult or easy a given system was to use under casual conditions. These judgments unfortunately tend to become very unreliable when different systems are judged under different circumstances or at different times.

- Various systems, mainly vocoders, produce a voice that is unnatural, either lacking speaker recognition or containing so much electrical accent that they sound

1. W.D. Voiers, A.D. Sharpley, and C.J. Hehmsoth, "Research on Diagnostic Evaluation of Speech Intelligibility," Final Report, Prepared by Tracor, Inc. for Air Force Cambridge Research Laboratories, Jan. 24, 1973.

2. J.C. Webster, "Compendium of Speech Testing Material and Typical Noise Spectra for use in Evaluating Communications Equipment," Technical Document 191, Naval Electronics Laboratory Center, Human Factors Technology Division, Sept. 13, 1972.

3. J.W. Preusse, "Consonant Recognition Test," Research and Development Technical Report ECOM-3207, Dec. 1969.

4. C.B. Grether and R.W. Stroh, "Subjective Evaluation of Differential Pulse Code Modulation Using the Speech 'Goodness' Rating Scale," 1972 IEEE Conference on Speech Communications and Processing, p. 175, Apr. 24-26, 1972.

5. "IEEE Recommended Practice for Speech Quality Measurements," *IEEE Transactions on Audio and Electroacoustics* AU-17, pp. 227-246, Sept. 1969.

6. B.J. McDermott, "Multidimensional Analysis of Circuit Quality Judgments," J. *Acoustical Soc. Amer.* 45 (No. 3), pp. 774-781, 1969.

7. W.P. Pachi, G.E. Urbanek, and E.H. Rothauser, "Preference Evaluation of a Large Set of Vocoded Speech Signals," *IEEE Transactions on Audio and Electroacoustics* AU-19 (No. 3), pp. 216-225 Sept. 1971.

similar to mechanical or robotlike voice reproductions. The present art of testing and rating these qualities has not been highly developed.

- Tests have been made which do not simulate transmission conditions so that they are representative of the actual operating environment. Since some systems degrade more than others under transmission conditions, it is possible for the ranking of systems to change substantially when tested under field conditions.* When comparing systems it is obviously important to test each system under comparable transmission conditions and also to make the best possible judgments about the importance of good performance in each given test environment.

METHOD OF QUALITY/ACCEPTABILITY RANKING

A method of formally making quality comparisons of various systems and environments has been developed at NRL and used for system comparisons under simulated transmission conditions. The test overcomes some of the objections mentioned and has proved useful in the absence of further research in the art of acceptability testing. A test of this type does not rate intelligibility; a system can sound good and be relatively unintelligible. Hence this quality/acceptability test should be supplemented by intelligibility testing.

The remainder of this report will be involved with a discussion of the NRL test procedure, its implementation, and typical test results.

TEST PROCEDURE

The test material comprises sentences processed through all possible pairs of the test systems (excluding the pairing of a system with itself), with a pair of systems in the reverse order being considered a second independent pair or test item. The number of test items is n(n-1), where n is the number of systems being tested. The test items are randomized by drawing slips from a hat without replacement. Thus it is virtually impossible to add additional systems to the test sequence once it has been run, without redoing the entire test. In the following typical tests, five systems were compared, resulting in a test requiring 20 items. For each new set of conditions (in this case, changing the dB//Hz ratio for all systems, where dB//Hz symbolizes the signal level in dB relative to noise in a 1-hertz bandwidth) a new randomization was generated.

The test material is generated with sentences from a source tape (Fig. 1) processed through all the test systems. A single sentence is used for each comparison, since it is difficult to make comparative judgments on different sentences. Although a different test sentence could be used for each test pair, this would greatly increase the labor of producing the test tape and is unnecessary, since intelligibility is not an issue for this test. Hence, for convenience the same sentence is used for all the comparisons. For extensive testing, more than one sentence might be desirable to reduce listeners boredom. In addition these tests have used only one talker for economy. In future tests either more talkers should be used

*For example, some systems degrade substantially when tested back-to-back, but not much more under adverse transmission conditions, whereas others degrade little back-to-back, but are substantially degraded with even slightly adverse transmission conditions.

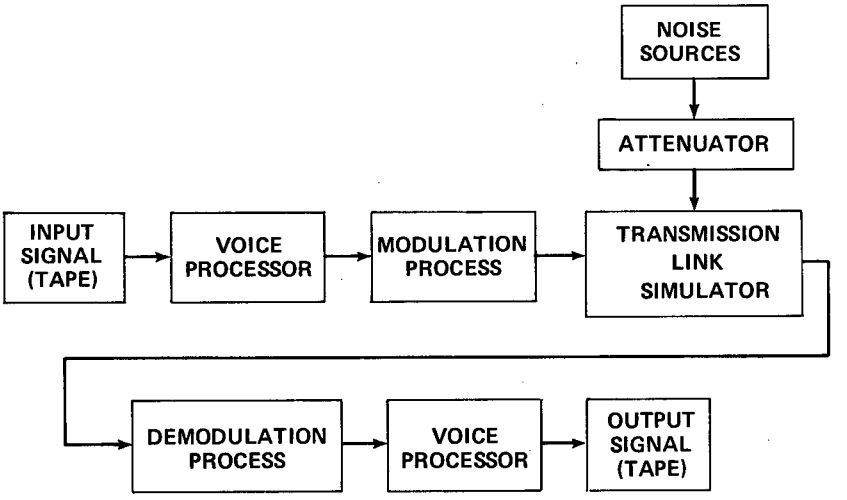


Fig. 1 — Setup to generate a sentence processed through all the test systems under simulated transmission conditions

or the impact of talker sensitivity would have to be rated independently of the preference test by one of the available multiple-talker-intelligibility tests.

For the tests to be reported, the test tapes were generated as follows: A loop of 6 seconds duration was prepared for a seven-track FM tape recorder (Fig. 2). The five system outputs using the same test sentence, requiring 3.5 seconds, were recorded approximately parallel on the tape using the loop splice as a synchronization marker. By use of a five-position switch, the appropriate test pairs were recorded as the loop rotated continuously. This provided approximately 2.5 seconds between the sentences, thereby allowing the listener to anticipate the next test sample. A mike switch was used in announcing item numbers, simultaneously disconnecting the loop playback between samples.

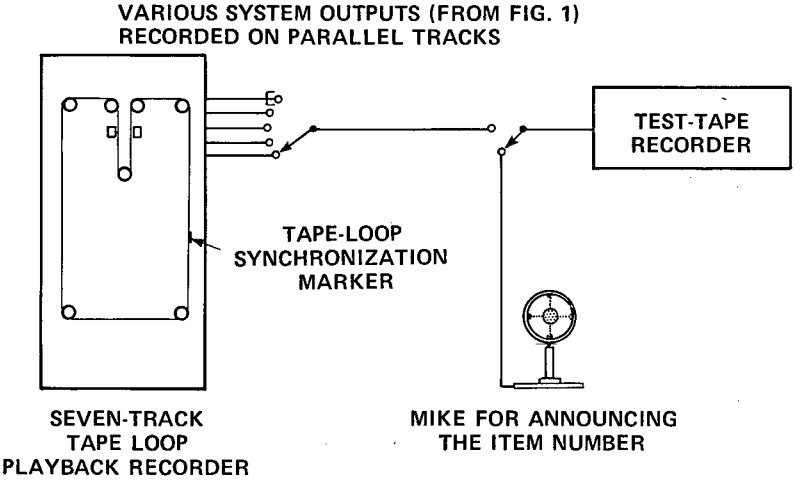


Fig. 2 — Setup to generate a test tape from recordings of the outputs in Fig. 1

The NRL test tapes are generally presented to groups of about six listeners at a time, using headphones. Telephone handsets could also be used in a quiet location. Use of loudspeakers is considered unsatisfactory, since the room acoustics at different seating locations and the directivity of the loudspeaker could produce an uncontrolled listener environment.

The listeners are instructed (Fig. 3) to vote for the system they prefer of each pair and to attempt to judge the system most preferred from a usage standpoint. Since the issue for military applications is acceptability and not fidelity of speech reproduction, no exposure to the original input test sentence is included. In taking the test the listeners record their responses on a response sheet (Fig. 4) which forces the responses to appear in the appropriate place for easier scoring. By applying master scoring masks to the response sheet, it is possible to quickly tabulate how many times each system is preferred by a given listener. The master scoring mask has cutouts which enclose the words "first" and "second" on the response sheet, thus ensuring proper registration of the two sheets in spite of paper size and margin variations. Cutouts also indicate the test being given. In the tests comparing five systems, each mask had eight cutouts corresponding to all the cases in which that system could be preferred (all the "correct" answers for that system). For a given system and a given listener the number of "correct" answers could vary from zero to eight. These data (and the equations given in Appendix A) were used to calculate the mean and standard deviation for each system. It was also valuable to know the number of listeners making a given score in order to plot a distribution which illustrated the response of the listeners. Table 1 shows how these data were tabulated.

Each sentence you will hear is transmitted over a voice communication system in a simulated transmission. You will be asked to say which one of two systems you prefer. After listening to both samples of a given test item, make a judgment as to which sample you prefer from the standpoint of quality (acceptability). Do not make your judgment on the basis of intelligibility as this is being done separately.

If you prefer the first of the two, check "a"; if you prefer the second, check "b." In making your judgment, include the fact that the noise background you hear with each sample would also be a characteristic of each system in use.

If you have difficulty in making a decision, give a best guess and move ahead. If there is no real preference, guesses will be random for all subjects and results will indicate a nonsignificant difference between the two (a tie).

If there are no questions, we will proceed.

Fig. 3 — Instructions given the quality-testing listeners

RESULTS OF A TYPICAL TEST

A typical data tabulation for the test group of 16 listeners from NRL is illustrated in Table 2a. From this table, it is possible to obtain the rank ordering of the test systems

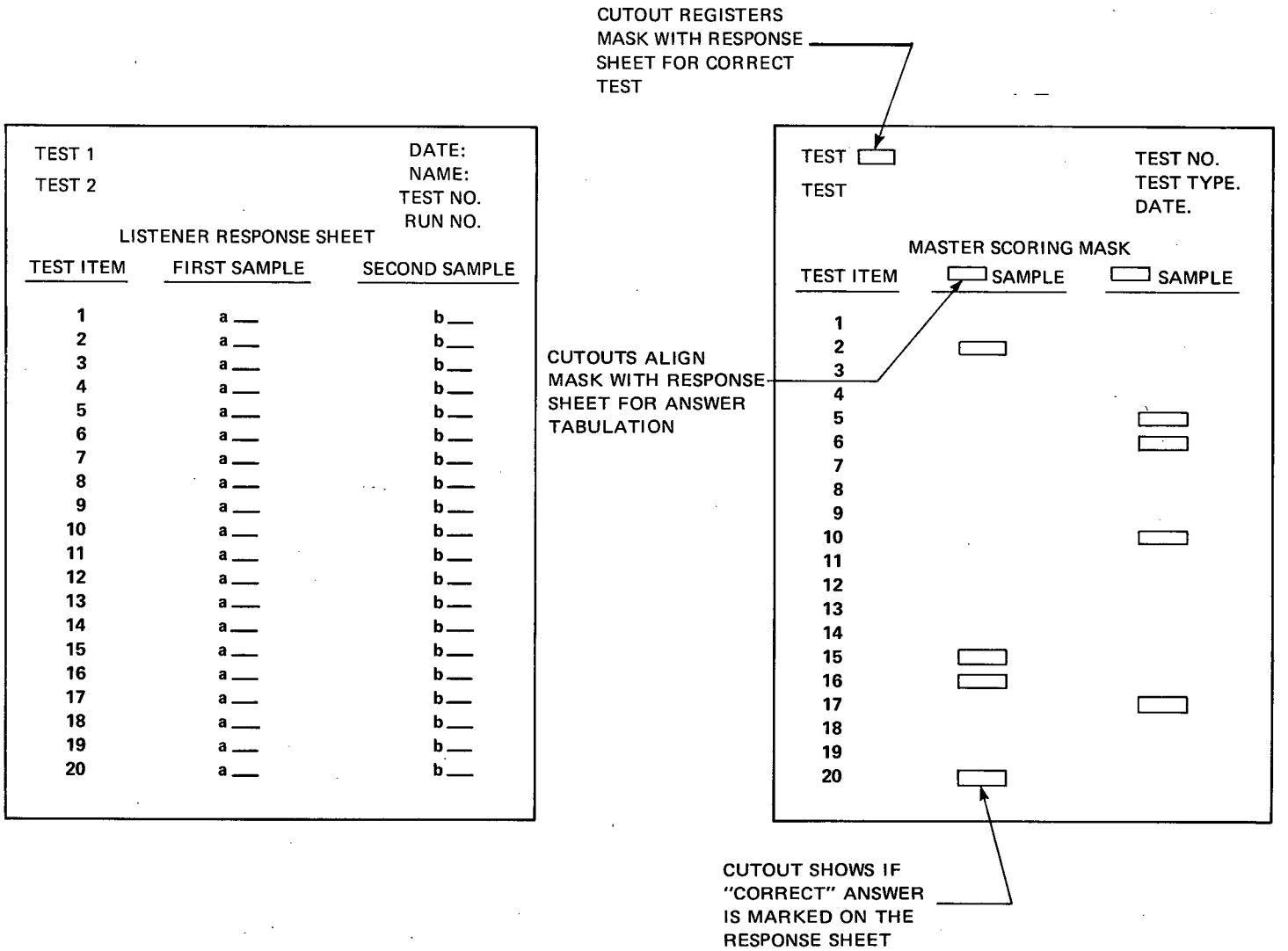


Fig. 4 — Listener-response sheet when five systems are compared and one of the five master scoring masks used in expediting the scoring of "correct" responses for each of the five systems. The eight cutouts correspond to the eight random pairings of the given system (pairings with each of the other four systems in alternative order).

Table 1
Form for Tabulating the Number of Listeners Who Scored the Various Possible Numbers of "Correct" Answers (Fig. 4) for Each of the Five Systems

Preferences Possible for Each System	Number of Listeners' Preferences Distributed Among Preferences Possible for Each System				
	Vocoder With Phase-Shift Keying at 2400 bps	Analog Voice Pulse-Duration Modulation With Phase-Shift Keying	Variable-Slope Delta Modulation With Phase-Shift Keying at 9600 bps	Analog Voice Frequency Modulation	Digital Voice Pulse-Duration Modulation
0					
1					
2					
3		Each entry in this table tallies the number of listeners who preferred a particular system a particular number of times			
4					
5					
6					
7					
8					

and the individual test system distributions. At this point in the data reduction a confidence interval at a 95-percent probability level is calculated. (The statement may then be made that there is a 95 percent probability that the true population mean will be within this confidence interval.) The confidence intervals are given in Table 2b. For these tests an overlap existed between the confidence intervals of some of the adjoining means in the rank ordering (Fig. 5). This indicated that these differences were not statistically significant for the sample size of 16 listeners, and that a larger sample would be required for accurate ranking of these cases. One reason for the large confidence-interval spread was believed to be the difficulty in deciding system preference when a judgment between different types of distortion had to be made. This could include cases in which distortion by the voice processor degraded the voice quality as opposed to either quantization noise from digitization or additive background noise.

For certain types of systems, listeners become opinionated, and the distribution for these particular systems appears to be non-Gaussian. This can also lead to wide confidence intervals and is of interest to explore as a means of insight into the nature of the listener reaction.

In the typical tests reported here, plots which demonstrate a varied listener reaction are shown in Fig. 6. Figure 6a is for a system in which the speech has a Gaussian noise background. There was a great deal of agreement among subjects about its relative acceptability. Figure 6b is for a system which sounds artificial. The apparent noise background

Table 2a
Tabulation That Resulted From a Typical Test of Five Systems With a Subject Population of 16 Listeners (44 dB//Hz)

Number of Preferences Possible For One System	Number of Listeners Scoring a Given Number of Preferences				
	Vocoder With Phase-Shift Keying at 2400 bps	Analog Voice Pulse-Duration Modulation With Phase-Shift Keying	Variable-Slope Delta Modulation With Phase-Shift Keying at 9600 bps	Analog Voice Frequency Modulation	Digital Voice Pulse-Duration Modulation
0	1	0	0	1	12
1	1	0	0	1	2
2	0	0	0	7	2
3	1	0	1	2	0
4	1	3	7	3	0
5	0	2	1	2	0
6	2	6	6	0	0
7	4	3	1	0	0
8	6	2	0	0	0

Table 2b
Mean, Standard Deviation, and 95% Confidence Interval for the Results Given in Table 2a

System	Mean	Std. Dev.	Confidence Interval
Vocoder with phase-shift keying at 2400 bps	6.00	2.53	4.66-7.34
Analog voice pulse-duration modulation with phase-shift keying	5.94	1.24	5.28-6.60
Variable-slope delta modulation with phase-shift keying at 9600 bps	4.94	1.13	4.34-5.54
Analog voice frequency modulation	2.69	1.35	1.97-3.41
Digital voice pulse-duration modulation	0.38	0.69	0.01-0.75

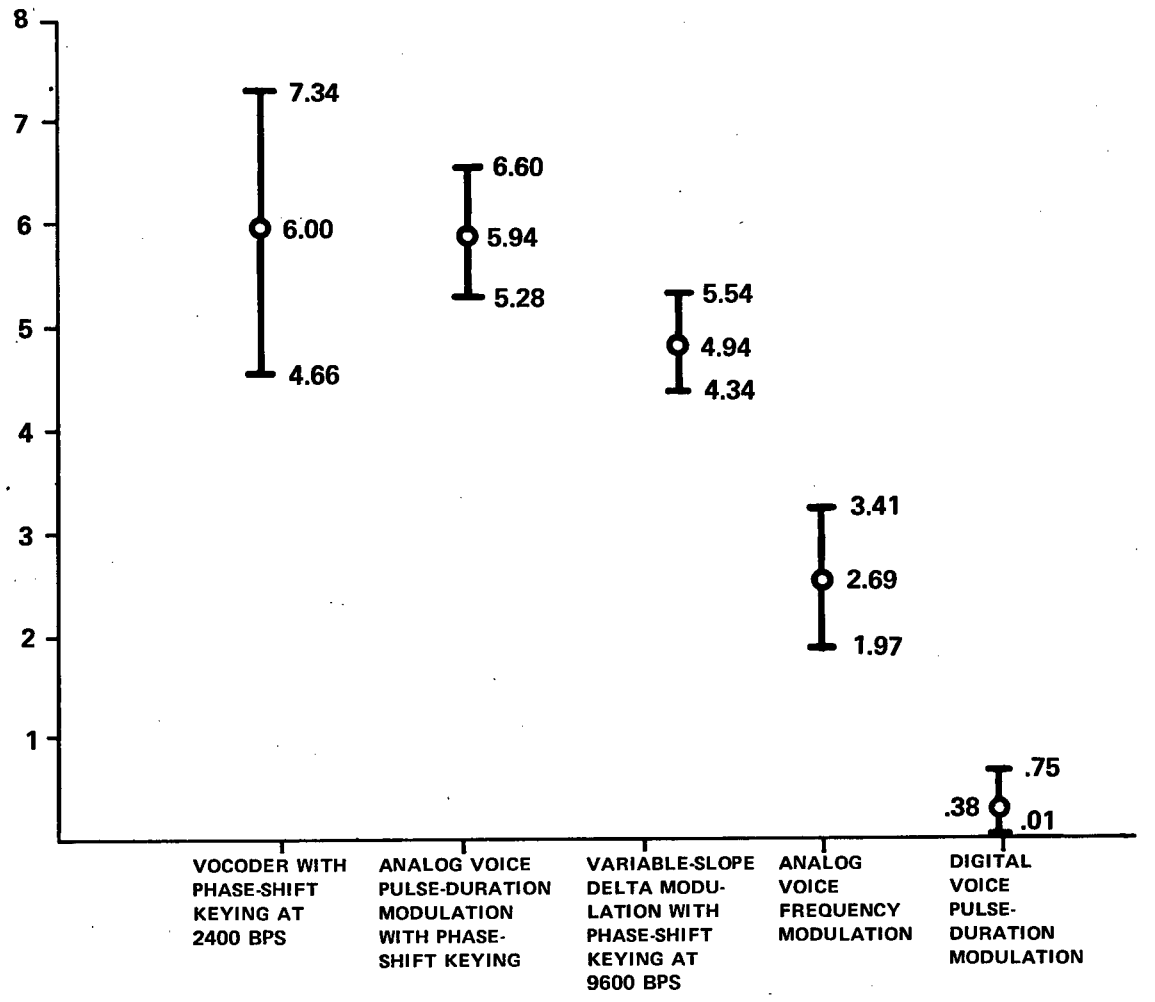
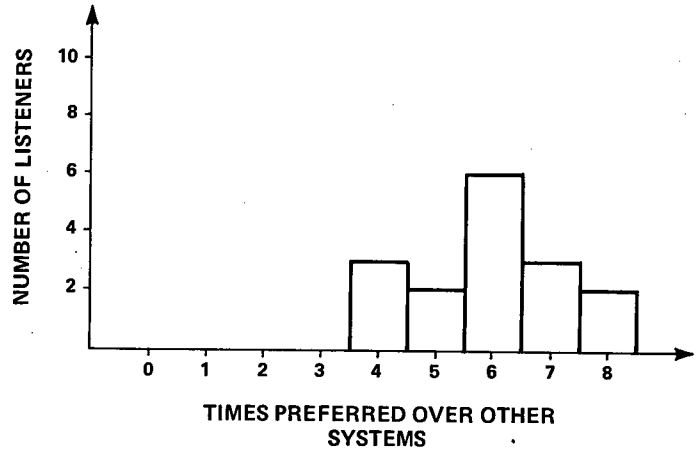


Fig. 5 — Rank ordering resulting from the quality and acceptability test of five systems by 16 listeners when the signal level relative to noise was 44 dB//Hz (Table 2b)

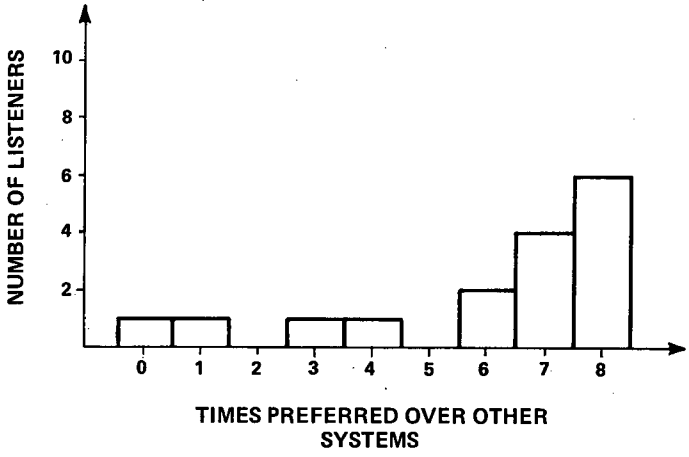
is low, but the listeners differed widely in their opinions as to the acceptability of the voice quality. This resulted in an unacceptably high standard deviation; hence a large sample N would be required to reduce the confidence interval of the mean, which decreases as the square root of N.

This test was rerun with a different group of listeners, and the system trends were the same. Listeners interviewed gave emotional responses to system quality consistent with their scores. It was evident that a wide divergency of opinion existed for the system of Fig. 6b, since one listener chose it zero times over other systems and other listeners chose the system all eight possible times over the others.

Since the tests using 16 listeners resulted in a less than completely satisfactory separation of the systems, additional listeners were scored using the same master tapes. These listeners included ten Navy Radiomen, seven attendees at a meeting of the Washington



(a) Analog voice pulse-duration modulation with phase-shift keying



(b) Vocoder with phase-shift keying at 2400 bps

Fig. 6 — Quality and acceptability distribution from data in Table 2a

Area Speech and Hearing Research Group, and fourteen subjects at the Defense Communications Agency Engineering Facility at Reston, Virginia. Though the taped tests were administered under slightly different conditions, the results were lumped together (Tables 3 and 4 and Fig. 7). The actual distributions for each system for two dB//Hz conditions are shown in Fig. 8. The additional number of subjects resulted in a noticeable narrowing of the confidence intervals but did not change the apparent rank ordering.

Table 3a
Tabulation That Resulted From the Same Test as in Table 2a But With a Subject Population of 47 Listeners (44 dB//Hz)

Number of Preferences Possible For One System	Number of Listeners Scoring a Given Number of Preferences				
	Vocoder With Phase-Shift Keying at 2400 bps	Analog Voice Pulse-Duration Modulation With Phase-Shift Keying	Variable-Slope Delta Modulation With Phase-Shift Keying at 9600 bps	Analog Voice Frequency Modulation	Digital Voice Pulse-Duration Modulation
0	1	0	1	1	32
1	1	0	0	4	10
2	1	1	3	19	4
3	4	1	5	11	0
4	4	7	13	6	1
5	0	10	7	5	0
6	6	16	13	1	0
7	8	8	5	0	0
8	22	4	0	0	0

Table 3b
Mean, Standard Deviation, and 95% Confidence Interval for the Results Given in Table 3a

System	Mean	Std. Dev.	Confidence Interval
Vocoder with phase-shift keying at 2400 bps	6.36	2.16	5.75-6.97
Analog voice pulse-duration modulation with phase-shift keying	5.68	1.33	5.30-6.06
Variable slope delta modulation with phase-shift keying at 9600 bps	4.70	1.56	4.26-5.14
Analog voice frequency modulation	2.77	1.27	2.41-3.13
Digital voice pulse-duration modulation	0.47	0.82	0.24-0.70

Table 4a
Tabulation That Resulted With the Same Subject Population as in Table 3a (47 Listeners) But For 50 dB//Hz

Number of Preferences Possible For One System	Number of Listeners Scoring a Given Number of Preferences				
	Vocoder With Phase-Shift Keying at 2400 bps	Analog Voice Pulse-Duration Modulation With Phase-Shift Keying	Variable Slope Delta Modulation With Phase-Shift Keying at 9600 bps	Analog Voice Frequency Modulation	Digital Voice Pulse-Duration Modulation
0	4	0	3	1	13
1	0	1	2	4	10
2	2	1	3	9	9
3	3	5	10	11	9
4	2	12	9	11	4
5	8	10	10	5	1
6	5	9	4	5	1
7	10	4	2	1	0
8	13	5	4	0	0

Table 4b
Mean, Standard Deviation, and 95% Confidence Interval for the Results Given in Table 4a

System	Mean	Std. Dev.	Confidence Interval
Vocoder with phase-shift keying at 2400 bps	5.64	2.14	5.03-6.25
Analog voice pulse-duration modulation with phase-shift keying	5.06	1.65	4.59-5.53
Variable slope delta modulation with phase-shift keying at 9600 bps	4.13	1.77	3.63-4.63
Analog voice frequency modulation	3.43	1.57	2.98-3.88
Digital voice pulse-duration modulation	1.75	1.52	1.32-2.28

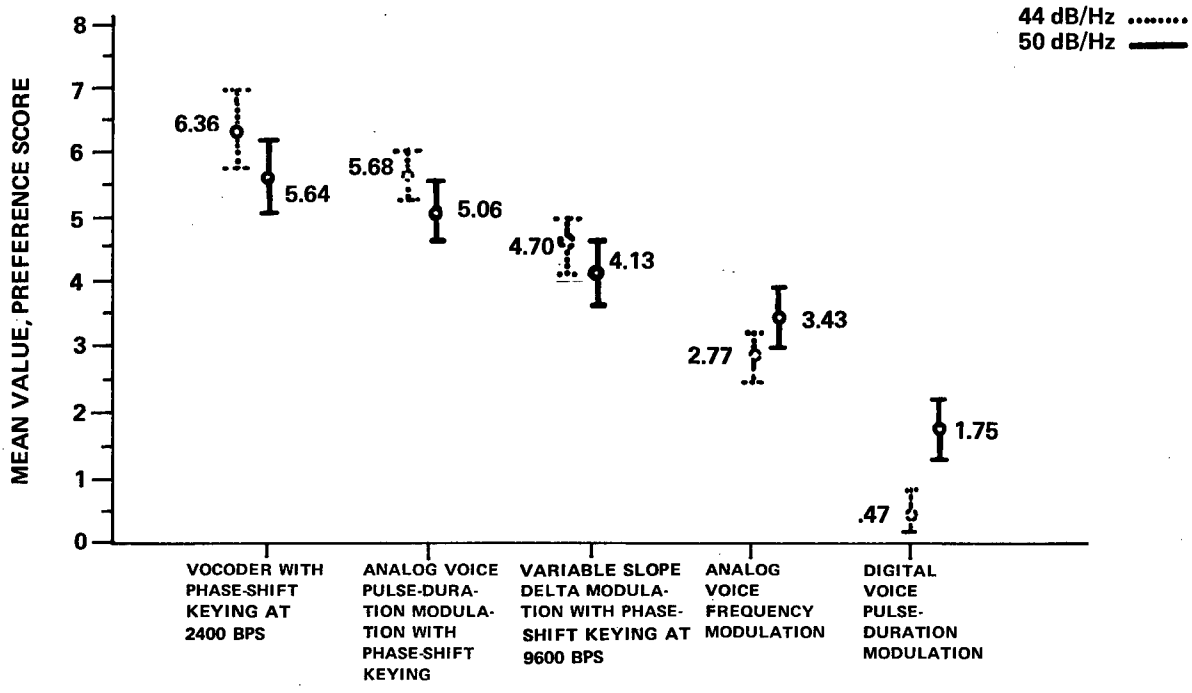


Fig. 7 — Rank ordering resulting from the quality-and-acceptability test of five systems by 47 listeners when the signal level relative to noise was 44 dB//Hz (Table 3b) and 50 dB//Hz (Table 4b)

One interesting observation based on the results as shown in Fig. 7 is that the ranking is much more separated in the 44-dB//Hz case than in the 50-dB//Hz case. This is apparently because some of the systems degraded much more under the influence of noise than others, thereby making the listener judgment easier. If a no-noise condition had been tested, then the rank ordering would probably have changed considerably, since the analog systems would improve much more than the vocoder or CVSD. This points up the necessity for carefully choosing realistic operating conditions for comparative tests in order for the results to be valid.

In regard to the larger spread of scores for the vocoder, various subjects were interviewed for their opinion of the vocoder versus the other systems. (The systems were identified to the listeners at this time.) Their opinions were consistent with their test scores. Though no sample of the original voice input was available, most listeners who disliked the vocoder complained about its lack of "high fidelity" and its unnatural quality.

COMMENTS ON THE TEST METHOD

Several positive points can be cited for using a test of this nature. The test is relatively easy to administer and score, and requires little instrumentation beyond the systems under test. It yields reliable data on the rank ordering of systems, though for cases like the vocoder the spread may be greater than desired, requiring a larger number of listeners to narrow the confidence interval.

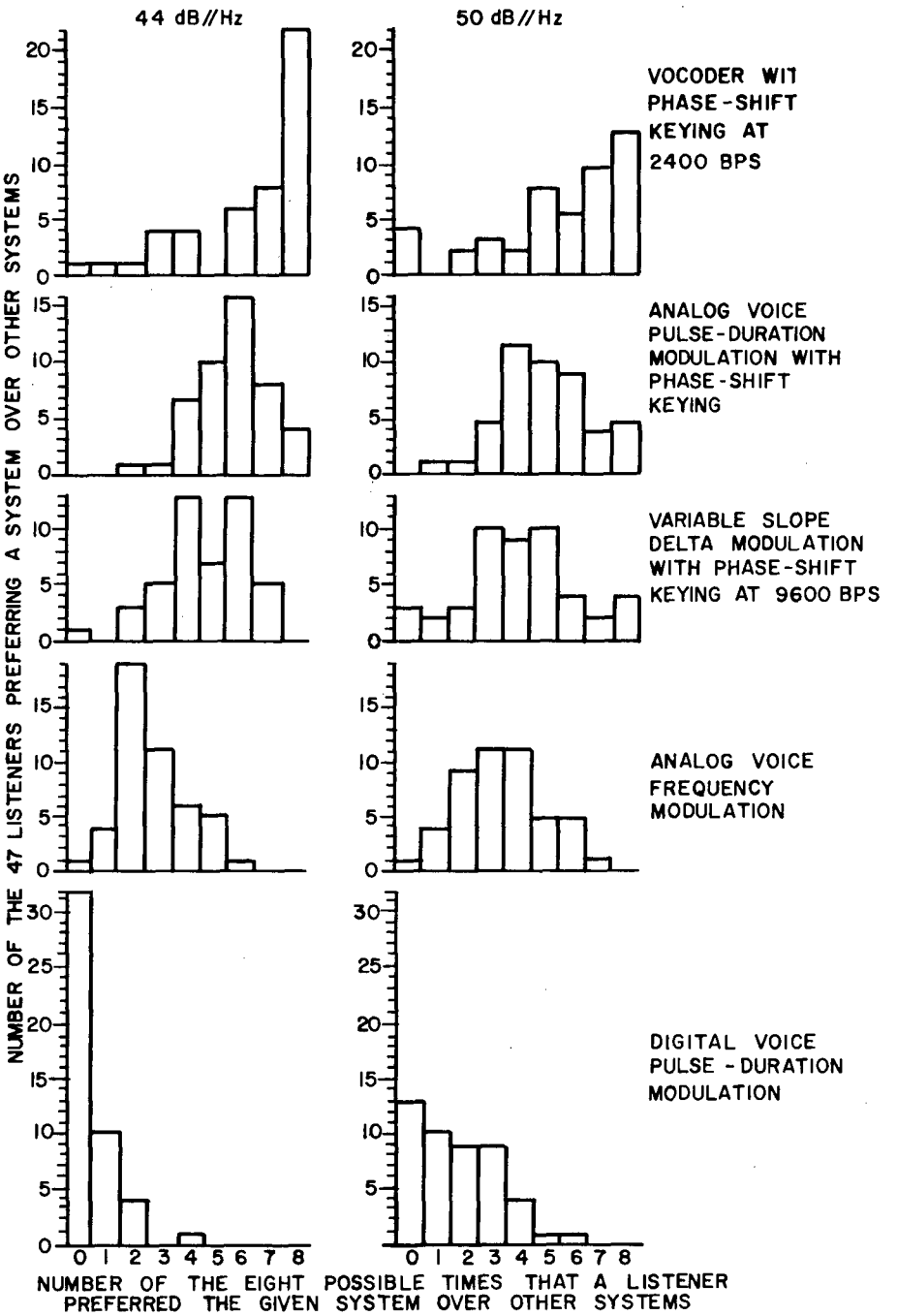


Fig. 8 — Quality and acceptability distributions given in Tables 3a and 4a

There are also drawbacks to this method. It is difficult to relate new systems to the old ranking without rerunning the entire test with inclusion of the new system. Another disadvantage is that the tape preparation and test administration becomes unduly cumbersome for more than about 10 systems due to the large number of test items.

CONCLUSIONS AND RECOMMENDATIONS

The problem of evaluating the quality and acceptability of narrowband voice digitzers is important as an aid to military procurement of such systems. It is desirable to test the systems under realistic operating conditions, because some systems degrade more than others in the presence of transmission noise. A method of rank ordering the systems has been described herein, and some typical results are given for a recent NRL test. It appears that this test method is satisfactory, provided a relative ranking is a satisfactory test result and no more than about 10 systems are to be tested. It would be desirable for a reliable method of quality and acceptability rating to be developed which would provide absolute scores rather than relative rankings so that new systems could be introduced at a later time without cumbersome modifications to the original test.

ACKNOWLEDGMENTS

The author acknowledges the support of this work under Irv Smietan, NAVELEX Code 3103, and the definition of the task and generous support of Don Himes, NRL Code 5420 Branch Head, and Lee Kline, NRL Code 5426 Section Head. He also thanks Frank Gentges for assistance in developing the test and preparing the test tapes, his wife Dr. W. R. Coulter (Psychologist) for suggestions regarding test makeup and procedure, and George Kang for assistance in making the initial calculations.

He also expresses his appreciation to Chief White, Chaltenham Naval Communication Center, to Major Leon Lake, DCA, and to Prof. J. M. Pickett, Gallaudet College, for use of subjects and facilities to improve confidence intervals.

He expresses special appreciation to Marc Rubenstein and the staff at Booz, Allen Applied Research for aid in making the final computations and in the preparation of this report.

APPENDIX A
GENERALIZED EQUATIONS FOR THE MEAN, STANDARD DEVIATION, AND CONFIDENCE INTERVAL FOR THE TEST RESULTS

MEAN

The mean value is

$$M = \sum_{i=0}^{2(n-1)} T_i P(T_i) ,$$

where n is the number of systems being compared, T_i is one possible test score, and

$$P(T_i) = \frac{S_i}{S_T} ,$$

in which S_i is the number of subjects in the test group who achieved the test score T_i and S_T is the total number of subjects in the test group.

DERIVATION OF THE STANDARD DEVIATION

The standard deviation is

$$\text{SD} = \sqrt{\text{variance}} = \sqrt{V^2} ,$$

where

$$V^2 = \sum_{i=0}^{2(n-1)} (T_i - M)^2 P(T_i) .$$

Expanding,

$$V^2 = \sum_{i=0}^{2(n-1)} (T_i^2 - 2MT_i + M^2) P(T_i)$$

$$= \sum_{i=0}^{2(n-1)} T_i^2 P(T_i) - \sum_{i=0}^{2(n-1)} 2MT_i P(T_i) + \sum_{i=0}^{2(n-1)} M^2 P(T_i) .$$

Since

$$\sum_{i=0}^{2(n-1)} P(T_i) = 1,$$

then

$$V^2 = \sum_{i=0}^{2(n-1)} T_i^2 P(T_i) - 2M \sum_{i=0}^{2(n-1)} T_i P(T_i) + M^2.$$

Further, since

$$\sum_{i=0}^{2(n-1)} T_i P(T_i) = M,$$

then

$$V^2 = \sum_{i=0}^{2(n-1)} T_i^2 P(T_i) - 2M^2 + M^2$$

$$= \sum_{i=0}^{2(n-1)} T_i^2 P(T_i) - M^2 .$$

Therefore

$$\mathrm{SD} = \sqrt{\sum_{i=0}^{2(n-1)} T_i^2 P(T_i) - M^2} .$$

CONFIDENCE INTERVAL

The confidence interval about the mean is

$$CI = M \pm t\left(\frac{SD}{\sqrt{S_T}}\right),$$

where the constant t determines the percent probability that the true mean is within the confidence interval:

t = Constant
= 2.37 for 7 subjects for a 95% CI
= 2.23 for 10 subjects for a 95% CI
= 2.15 for 14 subjects for a 95% CI
= 2.12 for 16 subjects for a 95% CI
= 1.96 for 50 subjects for a 95% CI.